

DOCUMENT RESUME

ED 187 745

TH 800 246

AUTHOR Hunyh, Hunyh; Saunders, Joseph C.
TITLE Accuracy of Two Procedures for Estimating Reliability of Mastery Tests. Research Memorandum 79-1.
INSTITUTION South Carolina Univ., Columbia.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Feb 79
GRANT NIE-G-78-0087
NOTE 14p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Kiawah Island, SC, February 22-24, 1979).
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; Error of Measurement; *Mastery Tests; Reliability; Statistical Bias; *Test Reliability; Test Theory
IDENTIFIERS Kappa Coefficient (Cohen)

ABSTRACT

Comparisons were made among various methods of estimating the reliability of pass-fail decisions based on mastery tests. The reliability indices that are considered are p , the proportion of agreements between two estimates, and $kappa$, the proportion of agreements corrected for chance. Estimates of these two indices were made on the basis of repeated or alternate test administrations and also on the basis of a single administration with the assumption of a beta-binomial or negative-hypergeometric distribution. Across a variety of test score distributions, test lengths, and mastery (cutoff) scores, the beta-binomial estimates tended to underestimate the corresponding population values. The percent of bias was small (about 2.5%) for p and somewhat larger (about 10%) for $kappa$. Both beta-binomial estimates had standard errors about one-half the size of the standard errors of estimates based on repeated test administrations. Though the beta-binomial estimates presumed equality of item difficulty, the data presented indicated that even gross departures from equality of item difficulty did not affect the amount of bias of the estimates. (Author/CTM)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

PUBLICATION SERIES IN MASTERY TESTING
University of South Carolina
College of Education
Columbia, South Carolina 29208

Research Memorandum 79-1
February, 1979

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

H. Huynh

ACCURACY OF TWO PROCEDURES FOR ESTIMATING
RELIABILITY OF MASTERY TESTS

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Huynh Huynh
Joseph C. Saunders III

University of South Carolina

Presented at the annual conference of the Eastern Educational
Research Association, Kiawah Island, South Carolina, February 22-24,
1979.

ABSTRACT

Single administration (beta-binomial) estimates for the raw agreement index p and the corrected-for-chance kappa index in mastery testing are compared with those based on repeated test administrations in terms of estimation bias and sampling variability. Across a variety of test score distributions, test lengths, and mastery (cutoff) scores, the beta-binomial estimates tend to underestimate the corresponding population values. The percent of bias is small (about 2.5%) for p and somewhat larger (about 10%) for kappa. Both beta-binomial estimates have standard errors about one-half the size of the standard errors of estimates based on repeated test administrations. Though the beta-binomial estimates presume equality of item difficulty, the data presented indicate that even gross departures from equality of item difficulty do not affect the amount of bias of the estimates.

This work was performed pursuant to Grant No. NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator.

ED187745

TM 800246

1. INTRODUCTION

In mastery testing reliability is often viewed as the consistency of mastery-nonmastery decisions across repeated test administrations (Huynh, 1976, 1978a; Subkoviak, 1976). Two reliability indices have been proposed and studied for mastery tests. They are the raw agreement index p and the corrected-for-chance kappa index (κ). The first index represents the proportion of examinees consistently classified in the same (mastery or nonmastery) category over two test administrations using the same form or two equivalent forms. It is assumed, of course, that the first testing does not induce any lasting change in the examinees. The second index, kappa, is defined as $\kappa = (p - p_c) / (1 - p_c)$, where p_c is the proportion of consistent classification expected under complete random assignment. Thus kappa reflects the extent to which test scores will improve the consistency of decisions beyond the level expected by random classification. The relationship between kappa and other parameters such as cutoff score and classical test reliability may be found in Huynh (1978a).

The definitions of both p and kappa assume the feasibility of repeated test administrations. This may not be practical in many instances. Under some conditions, p and kappa may be approximated from a single test administration. There are at least two procedures to accomplish this, namely, those described in Huynh (1976) and Subkoviak (1976). The Huynh procedure assumes that the test scores are distributed as predicted by a univariate or bivariate beta-binomial model. On the other hand, the Subkoviak technique, in its simplest form, assumes that test scores are distributed as predicted by a binomial distribution and that the regression of true score on observed test score is linear.

Subkoviak (1978) has provided a comparison of these two procedures using simulations with fifty repetitions. The data reported in Table 2 of his paper clearly indicate that both procedures act almost identically in terms of estimation bias and standard error. This is an expected result. Linear regression of

true score on observed score in the binomial error model automatically implies that the test score distribution under study must belong to the negative hypergeometric (beta-binomial) family (Lord & Novick, 1968, p. 316). Hence it appears that the conditions underlying the Subkoviak procedure are those of the beta-binomial distribution assumed in Huynh's paper (1976). For this reason and for inherent complexities in formulating inferential techniques associated with the Subkoviak procedure, this paper will be restricted to the beta-binomial model in the estimation of reliability for mastery tests.

The purpose of this paper is to compare the accuracy of two procedures for estimating reliability of decisions in mastery testing. One procedure is based on two test administrations; the other procedure relies on only one test administration and performs all computations assuming the appropriateness of the beta-binomial model for the test data under study. Sections 2, 3, and 4 deal with the asymptotic (large sample of examinees) nature of the estimates. Section 5 reports a simulation study for the case of small samples.

2. ASYMPTOTIC BIAS AND STANDARD ERRORS

Though the number of classification categories may be arbitrary, we will consider only the case of two categories, labeled mastery and nonmastery. The lowest score for which an examinee will be classified as a master will be referred to as the mastery (or passing) score in subsequent discussion.

First let us consider estimating p and κ by testing a sample of m examinees twice. Let p_{ij} be the proportion of examinees classified in the i -th category on the first testing and in the j -th category in the second testing. Here let $i = 0$ for a nonmaster and $i = 1$ for a master. Let the dot (.) bear the regular summation meaning. For example, the marginal proportion of masters on the first testing is $p_{1.} = p_{10} + p_{11}$.

The observed proportion* of consistent classifications in the sample at hand is $\hat{p}_R = p_{00} + p_{11}$ and the kappa index for this sample is

*The subscript R means repeated testings.

$$\hat{\kappa}_R = (\hat{p}_R - \hat{p}_C) / (1 - \hat{p}_C) \quad (1)$$

where $\hat{p}_C = p_{0.} p_{.0} + p_{1.} p_{.1}$. Under random sampling, $\hat{p}_R$ is an efficient statistic for the parameter p (Hogg & Craig, 1970, p. 372). In other words, $\hat{p}_R$ is unbiased and its standard error is equal to the Rao-Cramér lower bound. This standard error is $[p(1-p)/m]^{1/2}$. It may also be noted that $\hat{p}_R$ is also the maximum likelihood (ML) estimate of the population value of p and that $\hat{\kappa}_R$ is an ML estimate of the population value of κ . Its asymptotic (large sample) properties are well known. For example, $\hat{\kappa}_R$ follows an approximate normal distribution with mean κ and with a variance of

$$\frac{1}{m} \left[\frac{p(1-p)}{(1-p_C)^2} + \frac{2(1-p)(2pp_C - a)}{(1-p_C)^3} + \frac{(1-p)^2(b - 4p_C^2)}{(1-p_C)^4} \right] \quad (2)$$

where

$$a = p_{00}(p_{0.} + p_{.0}) + p_{11}(p_{1.} + p_{.1}) \quad (3)$$

and

$$b = \sum_{i,j} p_{ij}(p_{j.} + p_{.i})^2 \quad (4)$$

(Bishop et al., 1974, p. 396). In these formulae, all quantities listed are population values. When sample proportions are used in (2), the resulting value is an estimate for the variance of $\hat{\kappa}_R$. Finally, since the asymptotic mean of $\hat{\kappa}_R$ is κ , $\hat{\kappa}_R$ is asymptotically an unbiased estimate for this parameter.

Consider now estimating p and κ from a single test administration. The estimates*, $\hat{p}_B$ and $\hat{\kappa}_B$, are described in detail in Huynh (1976); the asymptotic standard errors of both estimates may be obtained via the formulae, tables, or computer program described elsewhere (Huynh, 1978b). In the latter paper it is also shown that $\hat{p}_B$ and $\hat{\kappa}_B$ are asymptotically unbiased estimates of p and κ .

3. A COMPARISON OF THE ASYMPTOTIC STANDARD ERRORS OF ESTIMATE FOR BETA-BINOMIAL TEST DATA

Whether estimation is based on repeated or single testings, $\sqrt{m}$ times the standard error (S.E.) of the estimate is (or is

*The subscript B refers to the beta-binomial model.

asymptotically) not a function of the sample size n . Thus n is not a significant factor in any comparison of the estimates as long as sufficiently large samples are to be considered. In this section and most subsequent ones, only the quantity $G = \sqrt{n} \times \text{S.E.}$ will be considered.

The comparisons described in this section are limited to test score distributions that follow the beta-binomial distribution. Strictly speaking, the procedure for estimating from a single administration (Huynh, 1976) is formulated only for this type of data.

The comparison was made for selected situations with $n = 5, 10, 20,$ and 30 test items. The test mean (μ) and KR21 reliability (α_{21}) were chosen such that the resulting test score distribution would be one of the following types: (i) U-shaped with the higher-density mode at the upper end of the score range, (ii) symmetric, (iii) unimodal with a mode somewhere between μ and n , or (iv) J-shaped. The passing score c was chosen such that the ratio c/n would be 60, 70, or 80%. The G -values for $\hat{\kappa}_R$ were computed via Equations (2), (3), and (4) with the p_{ij} proportions generated by the bivariate beta-binomial model. The G -values for $\hat{p}_B$ and $\hat{\kappa}_B$ were obtained via the computer program described in Huynh (1978b).

Table 1 reports the obtained G -values when the two procedures for estimating p and κ are used. The G -values in the table clearly demonstrate that the standard error associated with the single administration (beta-binomial) procedure is uniformly smaller than that encountered with the procedure using two test administrations. Over the thirteen situations reported in Table 1, the standard errors for the single administration procedure average 59.3% of those from repeated administrations for the p index and 53.2% for the kappa index.

4. A COMPARISON OF THE ASYMPTOTIC BIAS AND STANDARD ERRORS OF ESTIMATE FOR CTBS TEST DATA

This phase of the study is motivated by the fact that real test data rarely conform exactly to a well-specified model such as

TABLE 1

G-Values for Beta-Binomial Test Data

α	β	Shape	n	μ	σ	c	Index p			Kappa		
							p	$G(\hat{p}_B)$	$G(\hat{p}_R)$	κ	$G(\hat{\kappa}_B)$	$G(\hat{\kappa}_R)$
5.0	3.0	Unimodal	5	3.125	1.301	3	.687	.320	.464	.270	.763	1.021
						4	.645	.350	.479	.273	.752	.970
2.0	.5	J-Shaped	5	4.000	1.309	3	.872	.168	.334	.492	.713	1.226
						4	.811	.265	.391	.526	.619	.953
.5	.2	U-Shaped	5	3.571	1.850	3	.907	.145	.291	.765	.379	.727
6.0	6.0	Symmetric	5	2.500	1.279	3	.605	.412	.489	.210	.823	.978
10.0	5.0	Unimodal	10	6.667	1.863	7	.644	.331	.479	.277	.663	.966
						8	.661	.280	.473	.262	.660	.966
8.0	2.0	Unimodal	10	8.000	1.706	7	.799	.222	.401	.332	.677	1.175
						8	.714	.295	.452	.357	.630	.984
4.5	.5	J-Shaped	10	9.000	1.500	7	.921	.135	.269	.454	.785	1.637
12.0	8.0	Unimodal	20	12.000	3.024	12	.678	.269	.467	.342	.550	.949
						14	.704	.235	.456	.326	.561	.998
12.0	3.0	Unimodal	20	16.000	2.646	12	.918	.169	.275	.304	.677	1.796
						14	.821	.192	.383	.370	.591	1.201
3.0	.5	J-Shaped	20	17.143	3.576	12	.940	.087	.237	.637	.478	1.369
16.0	14.0	Unimodal	30	16.000	3.801	20	.787	.212	.409	.290	.585	1.178
						24	.964	.123	.185	.142	.557	2.448
18.0	2.0	Unimodal	30	27.000	2.535	20	.982	.081	.133	.246	.775	3.716
						24	.888	.169	.315	.373	.650	1.496
19.5	.5	J-Shaped	30	29.250	1.319	24	.990	.062	.099	.273	1.105	5.038

the beta-binomial distribution. It is based on a portion of the Comprehensive Tests of Basic Skills (CTBS) test data collected in the 1978 South Carolina Statewide Testing Program. Table 2 describes the various tests artificially assembled from CTBS subtests or from the entire battery. For each test in the listing, two alternate (hopefully equivalent) forms were created by pairing items on the basis of content and/or difficulty and randomly assigning the items in each pair to the alternate forms. For reasons which will be obvious later on, a number of tests were deliberately constructed of items of similar difficulty.

The number of items (n) was set at 5, 10, 15, and 20. The number of students, selected by taking every tenth case from the entire South Carolina file, ranged from $n = 1684$ to 6035. For each test, the value $D_{\max}$ represents the maximum discrepancy between the observed relative cumulative frequency and the corresponding expected frequency from the beta-binomial model. A significance level (P-value) of more than .20 indicates that the test data follow closely the beta-binomial distribution. On the other hand, P-values of less than .05 or .01 reveal substantial departures from the theoretical distribution.

For each test described in Table 2, the population values p_R , $G(\hat{p}_R)$, κ_R , and $G(\hat{\kappa}_R)$ were computed using the bivariate frequency distribution generated by the alternate forms. The corresponding parameters p_B , $G(\hat{p}_B)$, κ_B , and $G(\hat{\kappa}_B)$ were obtained by imposing the beta-binomial model on each of the two alternate forms and averaging the two sets of results. Now both $\hat{p}_B$ and $\hat{\kappa}_B$ are asymptotic unbiased estimates of p_B and κ_B (Huynh, 1978b). Also, since $\hat{p}_R$ is an unbiased estimate of p_R , and $\hat{\kappa}_B$ is an asymptotically unbiased estimate of κ_R , only the asymptotic bias of $\hat{p}_B$ and $\hat{\kappa}_B$ in estimating p_R and κ_R was explored. Thus, it follows that the percent asymptotic bias for $\hat{p}_B$ and $\hat{\kappa}_B$ is $100(p_B - p_R)/p_R$ and $100(\kappa_B - \kappa_R)/\kappa_R$, respectively. A negative bias indicates underestimation whereas a positive bias documents an overestimation. (We focused on p_R and κ_R because test reliability is typically approached from the standpoint of equivalent forms.) All computations reported in this section were carried out as in the previous section.

TABLE 2

Description of the CTBS Data Used in Sections 4 and 5

Case	n	M	diff	$D_{\max}$ ($\bar{X}$)	P-value	Grade	Description
5.1	5	1684	.056	1.80	>.20	3	Reading comprehension (paragraph)
5.2	5	1684	.107	0.68	>.20	3	Language expression
5.3	5	5543	.003	0.50	>.20	3	Total battery
10.1	10	1684	.060	2.24	>.20	3	Reading comprehension (sentences)
10.2	10	6035	.081	1.54	>.15	6	Reading vocabulary
10.3	10	5543	.007	2.02	<.05	3	Total battery
15.1	15	1684	.175	1.72	>.20	3	Science
15.4	15	1335	.022	3.85	<.05	6	Total battery
20.1	20	1684	.099	4.01	<.01	3	Mathematics
20.3	20	5543	.015	7.65	<.01	3	Total battery

Table 3 details the results of the various estimates for p_R and κ_R . The data indicate that the beta-binomial estimates ($\hat{p}_B$ and $\hat{\kappa}_B$) tend to underestimate the alternate-form population values. For the p index, the percent of bias ranges from -4.2 to 0.1 with an average of -2.3. A larger degree of bias, however, occurs in the estimation of kappa via $\hat{\kappa}_B$. The percent of bias for this estimate ranges from -17.5 to 0.9 with an average of about -7.8.

The larger bias of $\hat{\kappa}_B$ as compared with that of $\hat{p}_B$ is to be expected. With the factor $1 - p$ (which cannot exceed .50) in the denominator of Equation (1) defining kappa, the bias of $\hat{\kappa}_B$ is at least twice as large as that associated with $\hat{p}_B$. For situations in which a high proportion of examinees are to be classified either as masters or nonmasters, $1 - p_c$ is close to zero. As a consequence, the bias of $\hat{\kappa}_B$ will become more pronounced in those cases.

The beta-binomial model assumes that test items are equally difficult (Huynh, 1976). It would be natural to expect that the bias of the beta-binomial estimates would bear a positive (or direct) relationship with variation in item difficulty. This is not the case, however. The values of $D_{\max}$ in Table 2 clearly indicate that departures from the beta-binomial distribution show no resemblance to the standard deviation (σ_{diff}) of item difficulty.

TABLE 3

Percent Asymptotic Bias and G-Values for CTBS Test Data

Case	n	Cutoff Score	Index p			Kappa		
			$\bar{x}$ Bias	$G(\hat{p}_B)$	$G(\hat{p}_R)$	$\bar{x}$ Bias	$G(\hat{\kappa}_B)$	$G(\hat{\kappa}_R)$
5.1	5	3	-1.5	.174	.331	-7.1	.540	.403
		4	-3.5	.236	.350	-9.2	.485	.774
5.2	5	3	-2.6	.192	.348	-13.7	.664	1.064
		4	-4.7	.287	.391	-14.1	.593	.856
5.3	5	3	-2.8	.211	.364	-17.5	.734	1.148
		4	-3.4	.325	.429	-11.3	.667	.921
10.1	10	6	-2.9	.113	.256	-10.2	.329	.668
		8	-4.2	.147	.281	-9.7	.294	.604
10.2	10	6	-1.3	.136	.330	-5.3	.384	.832
		8	-3.6	.176	.347	-8.7	.345	.707
10.3	10	6	0.7	.136	.332	2.5	.537	1.165
		8	-1.2	.208	.385	-4.4	.441	.862
15.1	15	9	-2.6	.203	.403	-8.1	.407	.809
		13	-3.7	.164	.317	-7.6	.530	1.300
15.2	15	9	-1.9	.168	.393	-4.0	.351	.881
		13	-0.4	.141	.295	-7.1	.506	1.313
20.1	20	12	-2.7	.098	.241	-12.9	.412	1.040
		14	-2.8	.115	.292	-7.7	.353	.880
20.2	20	12	0.1	.132	.370	0.9	.267	.751
		14	-0.7	.121	.355	0.0	.283	.805

The same observation holds for the bias of $\hat{p}_B$ and $\hat{\kappa}_B$ as displayed in Table 3.

The G-values of Table 3 clearly show that the estimates based on the beta-binomial model have a smaller standard error of estimate than those based on alternate forms. Over all the situations considered, the standard error of $\hat{p}_B$ is about 50.4% of that of $\hat{p}_R$; the standard error of $\hat{\kappa}_B$ is about 50.2% of that of $\hat{\kappa}_R$. These results are consistent with those of Section 3.

5. A COMPARISON OF FINITE-SAMPLE BIAS AND STANDARD ERRORS OF ESTIMATE FOR CTBS TEST DATA

A simulation was conducted to study the sampling fluctuations of the estimates $\hat{p}_B$, $\hat{\kappa}_B$, and $\hat{\kappa}_R$ when sample sizes are of small or moderate size. This was done for samples of size $n = 20, 40$, and 60 . For each test, one thousand replications were used to obtain the observed percent of bias and G-value for $\hat{\kappa}_R$. As for estimates

based on the beta-binomial model, one thousand replications were simulated for each alternate form and the averages of the two sets of results were used to determine the bias and G-value for $\hat{p}_B$ and $\hat{\kappa}_B$.

Table 4 presents a summary of the results of simulation. The adequacy of the random number generator (more specifically, the IMSL (1977) subroutine GGUB) is documented by the near zero bias of $\hat{p}_R$ and the small fluctuation of the $G(\hat{p}_R)$ values for various sample sizes around the corresponding true values (enclosed in parentheses). The data reported in the table clearly show that, as in the case of large samples, the beta-binomial model tends to underestimate the parameters p_R and κ_R . The bias of $\hat{p}_B$ in estimating p_B averages -2.6%. For kappa, the bias of $\hat{\kappa}_B$ fluctuates around -11.0%. It is also interesting to note that the alternate form estimate, $\hat{\kappa}_R$, also tends to have a small negative bias.

TABLE 4

Percent Finite-Sample Bias and G-Values for CTBS Test Data

				$\hat{p}_B$		$\hat{p}_R$		$\hat{\kappa}_B$		$\hat{\kappa}_R$		
Case	n	Cutoff Score	m	$\bar{x}$		$\bar{x}$		$\bar{x}$		$\bar{x}$		
				Bias	$G(\hat{p}_B)$	Bias	$G(\hat{p}_R)$	Bias	$G(\hat{\kappa}_B)$	Bias	$G(\hat{\kappa}_R)$	
5.1	5	3	20	-0.5	.186	-.4	.325	-8.6	.617	+1.5	1.005	
			40	-0.1	.184	-.1	.335	-7.7	.569	-1.3	.936	
			60	-1.1	.188	-.1	.334	-7.4	.553	-0.3	.930	
			(Exact value		0	.331)						
10.1	10	7	20	-3.6	.141	-.1	.225	-11.9	.376	-1.3	.678	
			40	-3.9	.146	.2	.269	-11.6	.327	-1.2	.644	
			60	-4.0	.145	-.1	.268	-11.4	.304	-0.4	.625	
			(Exact value		0	.259)						
15.1	15	11	20	-3.4	.210	-.4	.395	-15.1	.543	-2.4	.949	
			40	-3.8	.206	.3	.402	-13.4	.525	-2.2	.927	
			60	-3.7	.203	-.2	.397	-13.0	.523	-0.1	.927	
			(Exact value		0	.392)						
20.1	20	14	20	-0.7	.141	-.2	.293	-12.7	.585	-5.0	1.017	
			40	-2.6	.137	0	.306	-10.2	.519	-1.3	.961	
			60	-2.6	.142	.2	.312	-9.2	.499	-2.2	.942	
			(Exact value		0	.292)						

The data in Table 4 show that the beta-binomial estimates have smaller sampling fluctuations than the alternate form estimates. For all situations reported in this table, the standard error of $\hat{p}_B$ is about 51.4% of that of $\hat{p}_R$; and the standard error of $\hat{\kappa}_B$ is about 56.9% of the standard error of $\hat{\kappa}_R$. These trends are very similar to those reported in the previous section.

6. DISCUSSION AND CONCLUSION

In this study the performance of a single administration estimate of reliability for mastery tests is compared with the behavior of the estimate based on two test administrations. The results clearly indicate that the single administration (beta-binomial) estimate for the raw agreement index p behaves very well. Not only does it show a negligible amount of negative bias, its sampling error is about half of that of the test-retest procedure. As for the kappa index, a moderate degree of negative bias (about ten percent) is displayed by the beta-binomial estimate. This estimate of kappa also has a standard error that is about one-half the corresponding value for the alternate form estimate. Though the beta-binomial estimates are originally derived for tests with items of equal difficulty, the data presented indicate that the bias of these estimates does not depend on the assumption of equal difficulty for test items. Our conclusion is that for testing situations involving tests like the CTBS (with items of a wide range of difficulty), the estimation for consistency of decisions in mastery tests may be safely carried out via one test administration with the beta-binomial model as a vehicle for computation.

BIBLIOGRAPHY

- Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, Massachusetts: The MIT Press.
- Hogg, R. V. & Craig, A. T. (1970). An introduction to mathematical statistics (Third edition). New York: The MacMillan Company.

- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement 13, 253-264.
- Huynh, H. (1978a). Reliability of multiple classification. Psychometrika 43, 317-325.
- Huynh, H. (1978b). Computation and inference for two reliability indices in mastery testing based on the beta-binomial model. Research Memorandum 78-1, Publication Series in Mastery Testing. University of South Carolina College of Education.
- IMSL Library 1 (1977). Houston: International Mathematical and Statistical Libraries.
- Lord, F. M. & Novick, M. R. (1963). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Company.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement 13, 265-276.
- Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability of mastery tests. Journal of Educational Measurement 15, 111-116.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance and comments of Anthony J. Nitko are gratefully acknowledged.